

What is claimed is:

1        1.     A system for finding compounds in a text corpus, comprising:  
2              a vocabulary comprising tokens extracted from a text corpus; and  
3              a compound finder iteratively identifying compounds having a plurality of  
4              lengths within the text corpus, each compound comprising a plurality of tokens,  
5              comprising:

6                  an *n*-gram counter evaluating a frequency of occurrence for one or  
7              more *n*-grams in the text corpus, each *n*-gram comprising tokens selected from the  
8              vocabulary; and

9                  a likelihood evaluator determining a likelihood of collocation for  
10              one or more of the *n*-grams having a same length, adding the *n*-grams having a  
11              highest likelihood as compounds to the vocabulary and rebuilding the vocabulary  
12              based on the added compounds.

1        2.     A system according to Claim 1, further comprising:  
2              an iterator selecting *n*-grams having a same length that is less than the  
3              same length as *n*-grams previously selected.

1        3.     A system according to Claim 1, wherein only some of the *n*-grams  
2              having a highest likelihood are added as compounds to the vocabulary.

1        4.     A system according to Claim 1, wherein the likelihood of  
2              collocation as a likelihood ratio  $\lambda$  is computed in accordance with the formula:

$$3 \quad \lambda = \frac{L(H_i)}{L(H_c)}$$

4        where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  
5         $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a  
6        pair of tokens.

1        5.     A system according to Claim 4, wherein the  $L(H_c)$  is determined,  
2              comprising dividing the *n*-gram into *n*-1 pairings of segments, calculating a

3 likelihood of collocation for each pairing of segments, and selecting the maximum  
4 likelihood of collocation of the pairings as  $L(H_c)$ .

1       6.     A method for finding compounds in a text corpus, comprising:  
2           building a vocabulary comprising tokens extracted from a text corpus; and  
3           iteratively identifying compounds having a plurality of lengths within the  
4       text corpus, each compound comprising a plurality of tokens, comprising:  
5               evaluating a frequency of occurrence for one or more  $n$ -grams in  
6       the text corpus, each  $n$ -gram comprising tokens selected from the vocabulary;  
7               determining a likelihood of collocation for one or more of the  $n$ -  
8       grams having a same length; and  
9               adding the  $n$ -grams having a highest likelihood as compounds to  
10      the vocabulary and rebuilding the vocabulary based on the added compounds.

1       7.     A method according to Claim 6, further comprising:  
2           selecting  $n$ -grams having a same length that is less than the same length as  
3        $n$ -grams previously selected.

1       8.     A method according to Claim 6, further comprising:  
2           adding only some of the  $n$ -grams having a highest likelihood as  
3       compounds to the vocabulary.

1       9.     A method according to Claim 6, further comprising:  
2           computing the likelihood of collocation as a likelihood ratio  $\lambda$  in  
3       accordance with the formula:

$$4 \quad \lambda = \frac{L(H_i)}{L(H_c)}$$

5       where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  
6        $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a  
7       pair of tokens.

1       10.    A method according to Claim 9, further comprising:

2           determining  $L(H_c)$ , comprising:  
3               dividing the  $n$ -gram into  $n-1$  pairings of segments;  
4               calculating a likelihood of collocation for each pairing of  
5    segments; and  
6               selecting the maximum likelihood of collocation of the pairings as  
7     $L(H_c)$ .

1           11.     A computer-readable storage medium holding code for performing  
2    the method according to Claim 6.

1           12.     An apparatus for finding compounds in a text corpus, comprising:  
2               means for building a vocabulary comprising tokens extracted from a text  
3    corpus; and  
4               means for iteratively identifying compounds having a plurality of lengths  
5    within the text corpus, each compound comprising a plurality of tokens,  
6    comprising:  
7               means for evaluating a frequency of occurrence for one or more  $n$ -  
8    grams in the text corpus, each  $n$ -gram comprising tokens selected from the  
9    vocabulary;  
10          means for determining a likelihood of collocation for one or more  
11   of the  $n$ -grams having a same length; and  
12          means for adding the  $n$ -grams having a highest likelihood as  
13   compounds to the vocabulary and means for rebuilding the vocabulary based on  
14   the added compounds.

1           13.     A system for identifying compounds through iterative analysis of  
2    measure of association, comprising:  
3               a stored limit on a number of tokens per compound; and  
4               a compound finder iteratively evaluating compounds within a text corpus,  
5    comprising:

6                   an *n*-gram counter determining a number of occurrences of one or  
7 more *n*-grams within the text corpus, each *n*-gram comprising up to a maximum  
8 number of tokens, which are each provided in a vocabulary for the text corpus;  
9                   a likelihood evaluator identifying at least one *n*-gram comprising a  
10 number of tokens equal to the limit based on the number of occurrences and  
11 determining a measure of association between the tokens in the identified *n*-gram  
12 and adding each identified *n*-gram with a sufficient measure of association to the  
13 vocabulary as a compound token, rebuilding the vocabulary based on the added  
14 compound tokens and adjusting the limit.

1                  14.       A system according to Claim 13, further comprising:  
2                   a stored upper limit on a number of identified *n*-grams; and  
3                   a limiter identifying a number of *n*-grams up to the upper limit based on  
4 the number of occurrences.

1                  15.       A system according to Claim 13, further comprising:  
2                   an iterator initially specifying the limit comprising a plurality of tokens  
3 per compound and subsequently decreasing the limit comprising a lesser plurality  
4 of tokens per compound.

1                  16.       A system according to Claim 13, wherein the measure of  
2 association between the tokens in the identified *n*-gram comprises a likelihood  
3 ratio  $\lambda$ .

1                  17.       A system according to Claim 16, wherein the likelihood ratio  $\lambda$  is  
2 calculated in accordance with the formula:

$$3 \quad \lambda = \frac{L(H_i)}{L(H_c)}$$

4                  where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  
5  $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a  
6 pair of tokens.

1           18.     A system according to Claim 17, wherein, for each pair of tokens,  
2      $t_1, t_2$ , in the identified  $n$ -gram, the independence hypothesis comprises  
3      $P(t_2 | t_1) = P(t_2 | \bar{t}_1)$  and the collocation hypothesis comprises  $P(t_2 | t_1) > P(t_2 | \bar{t}_1)$ .

1           19.     A system according to Claim 17, wherein the  $L(H_i)$  is computed  
2     for each pair of tokens,  $t_1, t_2$ , in the identified  $n$ -gram in accordance with the  
3     formula:

4           
$$\arg \max_{L(H_i)} \frac{L(t_1, t_2 \text{ form compound})}{L(n\text{-}gram \text{ does not form compound})}.$$

1           20.     A system according to Claim 13, further comprising:  
2         an initial vocabulary comprising a plurality of tokens extracted from the  
3     text corpus.

1           21.     A system according to Claim 20, further comprising:  
2         a parser parsing the tokens from the text corpus.

1           22.     A system according to Claim 13, further comprising:  
2         a filter determining the number of occurrences of one or more  $n$ -grams  
3     within the text corpus for only unique  $n$ -grams.

1           23.     A system according to Claim 13, wherein each text corpus  
2     comprises a plurality of documents comprising one of a Web page, a news  
3     message and text.

1           24.     A method for identifying compounds through iterative analysis of  
2     measure of association, comprising:  
3         specifying a limit on a number of tokens per compound; and  
4         iteratively evaluating compounds within a text corpus, comprising:  
5             determining a number of occurrences of one or more  $n$ -grams  
6     within the text corpus, each  $n$ -gram comprising up to a maximum number of  
7     tokens, which are each provided in a vocabulary for the text corpus;

8                   identifying at least one *n*-gram comprising a number of tokens  
9    equal to the limit based on the number of occurrences and determining a measure  
10   of association between the tokens in the identified *n*-gram; and  
11                   adding each identified *n*-gram with a sufficient measure of  
12   association to the vocabulary as a compound token, rebuilding the vocabulary  
13   based on the added compound tokens and adjusting the limit.

1       25.   A method according to Claim 24, further comprising:  
2       providing an upper limit on a number of identified *n*-grams; and  
3       identifying a number of *n*-grams up to the upper limit based on the number  
4    of occurrences.

1       26.   A method according to Claim 24, further comprising:  
2       initially specifying the limit comprising a plurality of tokens per  
3    compound; and  
4       subsequently decreasing the limit comprising a lesser plurality of tokens  
5    per compound.

1       27.   A method according to Claim 24, wherein the measure of  
2    association between the tokens in the identified *n*-gram comprises a likelihood  
3    ratio  $\lambda$ .

1       28.   A method according to Claim 27, further comprising:  
2       calculating the likelihood ratio  $\lambda$  in accordance with the formula:

$$3 \quad \lambda = \frac{L(H_i)}{L(H_c)}$$

4       where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  
5     $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a  
6    pair of tokens.

1           29.     A method according to Claim 28, wherein, for each pair of tokens,  
2      $t_1, t_2$ , in the identified  $n$ -gram, the independence hypothesis comprises  
3      $P(t_2 | t_1) = P(t_2 | \bar{t}_1)$  and the collocation hypothesis comprises  $P(t_2 | t_1) > P(t_2 | \bar{t}_1)$ .

1           30.     A method according to Claim 28, further comprising:  
2         computing the  $L(H_i)$  for each pair of tokens,  $t_1, t_2$ , in the identified  $n$ -gram  
3     in accordance with the formula:

4           
$$\arg \max_{L(H_i)} \frac{L(t_1, t_2 \text{ form compound})}{L(n\text{-}gram \text{ does not form compound})}.$$

1           31.     A method according to Claim 24, further comprising:  
2         constructing an initial vocabulary comprising a plurality of tokens  
3     extracted from the text corpus.

1           32.     A method according to Claim 31, further comprising:  
2         parsing the tokens from the text corpus.

1           33.     A method according to Claim 24, further comprising:  
2         determining the number of occurrences of one or more  $n$ -grams within the  
3     text corpus for only unique  $n$ -grams.

1           34.     A method according to Claim 24, wherein each text corpus  
2     comprises a plurality of documents comprising one of a Web page, a news  
3     message and text.

1           35.     A computer-readable storage medium holding code for performing  
2     the method according to Claim 24.

1           36.     An apparatus for identifying compounds through iterative analysis  
2     of measure of association, comprising:  
3         means for specifying a limit on a number of tokens per compound; and  
4         means for iteratively evaluating compounds within a text corpus,  
5     comprising:

6                   means for determining a number of occurrences of one or more *n*-  
7   grams within the text corpus, each *n*-gram comprising up to a maximum number  
8   of tokens, which are each provided in a vocabulary for the text corpus;  
9                   means for identifying at least one *n*-gram comprising a number of  
10   tokens equal to the limit based on the number of occurrences and means for  
11   determining a measure of association between the tokens in the identified *n*-gram;  
12   and  
13                   means for adding each identified *n*-gram with a sufficient measure  
14   of association to the vocabulary as a compound token, means for rebuilding the  
15   vocabulary based on the added compound tokens and means for adjusting the  
16   limit.